

ETHNICITY

Introduction

The ONS Longitudinal Study (LS) sample makes up about 1% of the total England and Wales population; it is the largest longitudinal study containing data on ethnicity in the UK. Therefore, it can be used to examine ethnic differences in many health and social outcomes, which are difficult to examine in other longitudinal studies owing to insufficient numbers.

A person's ethnicity is essentially a matter of self-identification. The 1991 Census was the first to ask about ethnicity¹, and the question was changed substantially for the 2001 and 2011 Censuses to reflect changes in social perception.

The aim of this guidance is to:

- Ascertain whether the Longitudinal Study (LS) is a suitable dataset for your ethnicity study.
- Specify the sample population for your study.
- Specify the variables that you will need to extract from the LS for your study.

To illustrate the types of ethnicity data that are available in the LS this guidance also works through an example research question. This will highlight many of the common issues that a researcher is likely to encounter. The research question that will be addressed is:

'What differences are there between ethnic groups with respect to the numbers of people who move from local authority housing to become owner-occupiers, and how have these numbers changed over time?'

¹ Ethnicity essentially refers to national, racial, cultural, religious, linguistic or other attributes that are perceived as characteristic of a distinct group of people.

Ethnicity data in the LS

Ethnicity in 1971

1971 Census (see 1971 [Census forms](#))

Although there are no questions that specifically ask about a person's ethnicity in the 1971 Census, questions B9 and B10 provide information on the LS member's country of birth and the country of birth of their parents. This information has been used to derive a variable in the LS called ETHOR7 (you can search for ETHOR7 in the [CeLSIUS Data Dictionary](#)). The LS also has information on people living in the same household as the LS member, therefore in cases where the LS member still lives with their parents, the parents' census records (in the non-member files) could be used to determine the country of birth of the LS member's grandparents. However, ETHOR7 has several limits:

- Its codes only identify the ethnicity of LS members from the New Commonwealth and Pakistan (NCWP)². Those LS members who did not originate from the NCWP were classified together as 'Not-NCWP'. Therefore, it may not always be safe to assume that 'Not-NCWP' means 'White British'.
- It cannot be reproduced for other people in the LS member's household, so for example, direct comparison between an LS member and his/her spouse is not possible.
- It groups Bangladeshis and Pakistanis together because Bangladesh was only formed as an independent nation in 1971. Therefore, respondents to the 1971 Census may have reported their country of birth as 'Pakistan' or 'East Pakistan', when they would later have reported 'Bangladesh'.
- The coding was only done for LS members who were 'traced' in 1971³. This was because the main area of analysis at the time of the 1971 Census was mortality, for which members had to be traced.

Understanding ETHOR7

ETHOR7 codes consist of three digits [see table 1 below]:

- Digit 1 – denotes whether an LS member's derived ethnicity is 'pure' or 'mixed' in reference only to the countries of birth of the member and their parents. 'Pure' is used for LS members born in the NCWP, and those born outside the NCWP whose parents were both born in the NCWP. 'Mixed' is used for those born outside the NCWP, with one parent born in the NCWP. All others are classified as 'non-NCWP'.
- Digit 2 – denotes the LS member's ethnic group.
- Digit 3 – represents the separate religious sub-groups for Asian members (based on analysis of surnames).

² 'New Commonwealth' means any Commonwealth country except Australia, New Zealand and Canada, and as such includes those countries from which the majority of immigrants to England and Wales originated during the 20th Century (Pakistan is mentioned separately because it left the Commonwealth in 1972 and re-joined in 1989).

³ Traced means that a record was found in the NHS central register.

Table 1: Meaning of digits in the ETHOR7 code

Number	First digit	Second digit	Third digit
0			Not Coded
1	Non-NCWP	Non-NCWP	Not applicable
2	Pure	West Indian	Hindu
3	Mixed	Indian	Muslim
4		Pakistani (incl Bangladeshi)	Sikh
5		African (excl African Asian)	
6		African Asian	
7		Mediterranean Commonwealth	
8		Remainder of Commonwealth	

Therefore, the ETHOR7 code 232 represents 'Pure', 'Indian', 'Hindu', while the code for a non-NCWP LS member would be 111 (i.e. 'non-NCWP', 'non-NCWP', 'Not applicable' because religion is only coded for Asian LS members).

[There is also a variable called ETHORX7 (see the [CeLSIUS Data Dictionary](#)) which corresponds to the second digit of the ETHOR7 code]

CeLSIUS staff have written STATA code to derive another variable from own and parents' country of birth, called ETHGRP7, which users can request as it may be preferred when ETHOR7 is not adequate.

Ethnicity in 1981 and 1991

1981 ethnicity (see 1981 [Census forms](#))

Compared to the 1971 Census, the 1981 Census has even less information that can be used to derive a person's ethnicity. Question 9 asks about a person's country of birth, but there is no question on parents' country of birth. Country of birth on its own is not sufficient to derive a reliable ethnicity variable, and it is only possible to see the parents' country of birth for those LS members who still live with their parents (through the non-member files). Some studies have however used the country of birth of the head of household as a proxy for the ethnicity of an LS member.

CeLSIUS staff have written STATA code to derive an ethnicity variable based on an LS member's country of birth (ETHGRP8), which users can request. It is compatible with ETHGRP7.

1991 ethnicity (see 1991 [Census forms](#))

The 1991 Census was the first census to include a question related to ethnicity (question 11 asks for the ethnic group). If we assume that ethnicity does not change over time, this means that it is possible to use the ethnicity from the 1991 Census to derive the ethnicity of an LS member in, for example, 1981. However, this will only work for those LS members who are present in both the 1981 and 1991 Censuses.

The responses to question 11 in the 1991 Census form appear in LS variable ETHNIC9 (see the [CeLSIUS Data Dictionary](#)). By comparing the values of ETHNIC9 with those under question 11 in the

1991 Census form, it can be seen that ETHNIC9 has many more categories. This is because on the census form under the tick boxes for 'Black-other' and 'Any other ethnic group', there is room for respondents to write in their own response, which led to the derivation of a further 28 categories on top of the seven main categories specified on the form. Table 2 presents the coding scheme for ETHNIC9, and shows that values above 70 represent these 'Other' categories.

Table 2: ETHNIC9 categories and codes

ETHNIC9 category	ETHNIC9 code
White	0
Black Caribbean	1
Black African	2
Indian	3
Pakistani	4
Bangladeshi	5
Chinese	6
Black-Other: British	70
Black-Other: Caribbean Island	71
Black-Other: North African	72
Black-Other: Other African	73
Black-Other: East African Asian	74
Black-Other: Indian Subcontinent	75
Black-Other: Other Asian	76
Black-Other: Other	77
Black-Other: Black/White	78
Black-Other: Asian/White	79
Black-Other: Other Mixed	80
Other-Other: British, Ethnic Group indicated	81
Other-Other: British, no Ethnic Group indicated	82
Other-Other: Caribbean Island	83
Other-Other: North African	84
Other-Other: Other African	85
Other-Other: East African Asian	86
Other-Other: Indian Subcontinent	87
Other-Other: Other Asian	88
Other-Other: Irish	89
Other-Other: Greek (including Greek Cypriot)	90
Other-Other: Turkish (including Turkish Cypriot)	91
Other-Other: Other European	92
Other-Other: Other answers	93
Other-Other: Black/White	94
Other-Other: Asian/White	95
Other-Other: Mixed White	96
Other-Other: Other mixed	97

Researchers need to decide what to do with these groups in terms of their study. For example, do they move 'Black-Other/Caribbean Island' into the 'Black-Caribbean' group, or is it better to keep the 'Black-Other' group together? Ultimately, this is a matter for the researcher to decide - there is no 'right answer'.

When deciding whether and how to re-categorise, consider the number of people in each of these groups. If a subgroup contains very few people, it may not be worth worrying about. Table 3 (see

Excel file) shows the number of people in the census, and in the LS, for each category of ETHNIC9. You can see that certain of the groups we discussed on the previous page are in fact very small and so recoding or excluding them would make little difference to most analyses.

Ethnicity in 1991 Census outputs

The system of re-categorisation used in published census reports is the 1991 ethnicity output classification. If you intend to compare your results from the LS with published census results, you will need to use this output classification. Table 4 below shows how the coding scheme for ETHNIC9 maps onto the categories of the 1991 ethnicity output classification. You will notice that the output classification introduces a new category, Other-Asian, which contains those subgroups of Black-Other and Other-Other that are specifically described as Asian.

Table 4: Ethnicity census output classification

Ethnicity census output category	ETHNIC9 codes
White	00, 89, 90, 91, 96
Black-Caribbean	01, 71, 83
Black-African	02, 73, 85
Indian	3
Pakistani	4
Bangladeshi	5
Chinese	6
Black-Other	70, 77, 78, 80
Other-Asian	74, 75, 76, 86, 87, 88
Other-Other	72, 79, 81, 82, 84, 93, 94, 95, 97

For the remainder of this guidance ethnicity in 1991 is categorised using the main ETHNIC9 categories only (White, Black-Caribbean, Black-African, Indian, Pakistani, Bangladeshi, Chinese, Black-Other, Other) (see table 2 above).

Comparing country of birth with reported ethnicity

Table 5 (see Excel file) shows 1981 country of birth and 1991 ethnic group for those LS members who are present at both Censuses. You can see that only 50% of LS members who in the 1991 Census said that their ethnic origin was Black-Caribbean were born in the Caribbean. Similarly, only 51% of LS members who in the 1991 Census said that their ethnic origin was Indian were born in India, and only 46% of those who said that their ethnic origin was Chinese were born in Hong Kong, China and Taiwan. The relevant cells have been shaded in the table.

We should note that not all LS members present in the 1981 Census were subsequently present in the 1991 Census: they might have died or emigrated in the intervening period; they might not have completed the census form in 1991; or their census return may not have been matched to their LS record. Therefore, 1991 ethnicity data will only be appropriately applied to 1981 Census respondents if a sufficiently large proportion of LS members were present in both 1981 and 1991. Table 6 (see associated Excel file) presents information on the number of LS members present in the 1981 and

1991 Censuses according to their country of birth from the 1981 Census. As can be seen, 21% of LS members present in the 1981 Census, were not present at the 1991 Census, and this was worse for LS members from certain countries. For example, 36% of LS members born in the Caribbean who were present at the 1981 Census were not present in the 1991 Census, which means that using 1991 ethnicity for this group would bias the results. Similarly, higher percentages of LS members who were present at the 1981 Census but not the 1991 Census were also seen for LS members who were born in Sub-Saharan Africa (31%), Pakistan (35%), Bangladesh (35%) and Hong Kong, China and Taiwan (37%).

Table 7 (see Excel file) below presents a cross-tabulation of ETHOR7 against ETHNIC9. In the table, ETHOR7 is reduced to its middle digit only, which specifies the ethnic group (called ETHORX7 in the [CeLSIUS Data Dictionary](#))⁴. ETHNIC9 has also been regrouped into its main categories (i.e. subgroups of Black-Other and Other-Other respectively have been merged). It can be seen that those LS members classified as 'White' in ETHNIC9 were similarly classified in ETHORX7 (98% of the ETHNIC9 'White' category were in the 'Non-NCWP' category for ETHORX7). In contrast, the ETHNIC9 category that showed the greatest level of ETHORX7 divergent classification was 'Pakistani' (16.3% were classified as 'Indian' according the ETHORX7 classification based on own and parents' country of birth).

Ethnicity in 2001

In the 2001 [Census forms](#) there is a more detailed question on ethnicity (question 8), and a new question on religion (question 10).

Question 8 asks '*What is your ethnic group?*' and contains 16 tick boxes with five write-in boxes. These are separated into five sections (A, B, C, D and E). Respondents were asked to choose one section and tick the box that was most appropriate. A new Mixed group is included in 2001, and the White category is split into British, Irish or Other White.

The LS contains two variables, ETHGRP0 and ETHP0 (see the [CeLSIUS Data Dictionary](#)):

- ETHGRP0 (Ethnic group 2001) uses the 16-category classification that corresponds to the 16 tick boxes on the census form.
- ETHP0 (Ethnicity in 2001) combines the information from the tick boxes and the write-in boxes in the ethnicity question and consists of more than 100 categories.

Values in the 2001 Census that were missing or inconsistent have been imputed. These imputed values have been flagged in the LS (ETHPIMP). The proportion of values for ethnic group that were imputed in the 2001 Census was 2.91%.

⁴ Non-NCWP, West Indian, Indian, Pakistani (including Bangladeshi), African, African-Asian, Mediterranean Commonwealth, Rest of Commonwealth, Not traced at 1971 Census.

The numbers of people in each ethnic group in the LS in 2001 are shown in table 8 (see Excel file).

Question 10 which asks '*What is your religion?*' was voluntary in England and Wales. There are seven tick boxes starting with 'None' and ending with a write-in box for 'Any other religion'. Since this question was voluntary, no values were imputed in the LS variable in RELP0 (Religion 2001), and so there is a 'Not stated' category. The numbers of people for each religion in the nine standard census categories in the LS at 2001 are presented in table 9 (see Excel file).

Ethnicity in 2011

In the 2011 [Census forms](#) there are questions on ethnicity (question 16) and religion (question 20), which are similar to the questions in the 2001 Census form.

Question 16 in the 2011 Census form has 18 tick boxes with five write-in boxes, which are again separated into five sections (A, B, C, D and E). An additional tick box was added to section A: White for 'Gypsy or Irish Traveller', and an additional tick box was added to section C: Asian/Asian British for 'Chinese'. Section E, which had been for 'Chinese or other ethnic group' in the 2001 Census form, was for 'Other ethnic group', with tick boxes for 'Arab' and 'Any other ethnic group'.

The LS contains two ethnicity variables ETHGRP11 and ETHP11 (see the [CeLSIUS Data Dictionary](#)):

- ETHGRP11 (Ethnic group 2011) uses the 18-category classification that is used on the census form.
- ETHP11 (Ethnicity in 2011) combines the information from the tick boxes and the write-in boxes in the ethnicity question, and again comprises more than 100 categories.

The numbers of people in each ethnic group in the LS in 2011 are shown in table 10 (see Excel file).

Question 20 in the 2011 Census form, the religion question, was again a voluntary question in England and Wales, and it had the same tick boxes as the 2001 religion question. Again, since the question was voluntary, missing values were not imputed, and therefore there is a 'Not stated' category. As in 2001, there are two religion variables in the LS for religion in 2011 (RELP11 and RELGP11). RELP11 is similar to RELP01 in that it included categories corresponding to the write-in box under the tick box for 'Any other religion'. RELGP11 groups religion into the nine standard census categories (i.e. those corresponding to the tick boxes with additional categories for 'No religion' and 'Religion not stated').

Consistency of ethnicity groups between 1991, 2001 and 2011

Comparing ethnicity at 1991, 2001 and 2011 is not straightforward because of the introduction of the 'Mixed' category in 2001, and the 'Arab' group (under section E) in the 2011 Census. The 1991 classification tends to be used when compatibility with country of birth in 1971 or 1981 is needed; but where a sample is drawn from 1991 or later, the 2001 or 2011 classification may be more useful.

However, not everyone enumerated in 1991 is also enumerated in 2001. Forward linkage rates show the proportion of people at one census who are present at a subsequent census, taking into account deaths and emigrations. Reports on the quality of linkage between each of the censuses have been produced and can be found on the Related downloads section of ONS' [Longitudinal Study \(LS\)](#) website. These show that people from minority ethnic groups present at the 1991 Census were less likely to be present at the 2001 Census. The forward linkage rate for the White group is 89.1%, but is only 73.7% for minority ethnic groups. Some groups have rates of less than 70%; Black Africans have a rate of 61.5%; Black Caribbean, 69.3%; and Other Asian, 67.9%⁵. For individuals present at the 2001 Census, again, those from the minority ethnic groups were less likely to be present at the 2011 Census. The forward linkage rate for British White was 89.2% but was <70% for Other White and Other Ethnic Group⁶.

Other ethnicity data

Event records for cancer (table CANC in the [CeLSIUS Data Dictionary](#)) also have information on ethnicity (ETHNICCC variable). This information comes from a question on ethnicity which is included on the cancer registration forms. However, this question was only included on cancer registration forms from 1993 onwards and has been poorly completed, and therefore it is not a useful source of ethnicity data for LS members.

⁵ Longitudinal Study Development Team, ONS (2001) Quality of Linkage between the 1991 and 2001 Censuses, ONS (Newport)

⁶ Longitudinal Study Development Team, ONS (2015) Quality of Linkage 2001 to 2011. ONS (Titchfield)

Example research question

'What differences are there between ethnic groups with respect to the numbers of people who move from local authority housing to become owner-occupiers, and how have these numbers changed over time?'

Housing tenure in the census

To examine this question, you need to identify LS members who were living in local authority housing at one census and were owner-occupiers at the next census. You can then compare these numbers with the total number of people who were in LA housing at the first census, for whom housing tenure records are available at the following census. This can be done for separate time-periods, e.g. 1971 to 1981, 1981 to 1991, and 1991 to 2001.

Housing tenure information is available for LS members in variables TENURE7 (1971), TENURE8 (1981), TENURE9 (1991), TENH0 (2001) and TENH11 (2011). Looking at the information on these variables in the Data Dictionary shows that although there is some consistency in the coding of these variables, there is also ambiguity. For example, it is possible to identify all owner-occupiers by combining different categories from each census (and deciding how to allocate shared ownership), but for individuals who are living in local authority housing there is ambiguity about the status of New Town housing. In TENURE7, there is a category for renting from the council or New Town Development Corporation. In TENURE8 however none of the categories mentions New Town Development Corporation (by looking at the 1981 Census form, you can see that the tenure question [question H2] includes council and New Town in the term 'local authority'). In TENURE9, there is a separate category for New Town Development Corporation, which also includes Housing Action Trusts. In TENH0 and TENH11 there is no mention of New Towns.

As has been shown above, the most helpful ethnicity information that we have is contained in the variables ETHNIC9 for 1991, ETHGRP0 for 2001 and ETHGRP11 for 2011. If we restrict the sample to LS members who were present in 1991 then the sample will be smaller than if we extend it by including LS members who were present in 1971; however, this would introduce problems of categorisation. Looking at table 7 (see associated Excel file) shows that some ETHOR7 categories can be used to denote some of the ethnic groups in ETHNIC9. Non-NCWP in ETHOR7 can be used to denote White and African-Asian can be used to denote Indian. Therefore, you could add LS members in 1971 who were in these groups to increase the sample size, by sacrificing some certainty about the attribution of ethnicity.

Table 11 (see Excel file) shows the number of LS members coded for ethnicity in 1971 and 1991 respectively, classified by ethnic group using the main ETHNIC9 classification, with ETHOR7 categories allocated to the nearest equivalent. For example, those classed as 'African-Asian' by ETHOR7 are put into the 'Indian' group, and the united Pakistani/Bangladeshi group from 1971 are all classed as 'Pakistani'. The table shows how the addition of extra people coded for ETHOR7 who were not coded for ETHNIC9 increases the sample size, without necessarily changing the results.

To complete the study, you need to know how the numbers of people moving from local authority housing to owner-occupation have changed over time, by producing tables for 1971-1981, 1981-1991, 1991-2001 and 2001-2011. For both the 1971-1981, 1981-1991 and 1991-2001 analyses you can use ETHNIC9 categories to examine ethnicity of sample members. For the 2001-2011 analyses, you can use ETHGRP0 or ETHGRP11 categories.

Tables 12-15 (see Excel file) show the percentage of LS members who moved from Local Authority housing to owner-occupation by ethnicity in the periods 1971-1981, 1981-1991, 1991-2001, and 2001-2011 respectively.

It is noticeable that the percentage of people moving from local authority housing to owner-occupation in the period 1981-1991 has increased overall when compared to the period 1971-1981 (from 26.8% to 41.4%). The same sort of change is observed in each of the ethnic groups except for the Indian and Pakistani groups where the percentages for 1981-1991 are similar to those for 1971-1981, and the Chinese group where the percentage for 1981-1991 is lower than that for 1971-1981.

For the period 1991-2001 and the period 2001-2011, the percentage of people moving from local authority housing to owner-occupation has decreased overall when compared to the period 1981-1991 (to 28.4% in 2001 and 22.4% in 2011), and this decrease is seen across the ethnic groups.

It is clear that, for many ethnic groups classified in 2001, the sample sizes quite low (table 15a), and the study would benefit from grouping of categories for this last period (table 15b).

Summary

In this guidance you have seen:

- The ethnicity information that is contained in the LS, and where it originates from.
- How this information is coded as variables in the LS.
- How ethnicity data can be used retrospectively or prospectively, and the issues that this raises.
- How the ethnicity data in the LS might be used in a real study and what sample sizes are possible.